Car Price Prediction Using Machine Learning

[1]Ankit Ranjan, [2]Ankit Kumar, [3]Ashwini Kumar, [4]Anurag Gupta
[1,2,3,4]*School of Computer Science and Engineering Galgotias University, Noida, India*

***Abstract*—The objective of this study was to assess how machine learning (ML) techniques can be used to predict vehicle prices. Vehicle price prediction is a complex process, as there are many variables that can influence a vehicle's market price. The automotive industry has continued to grow, and there are an ever-increasing number of variables that can influence an automotive vehicle's price, such as manufacturer and model, fuel economy, additional features, etc. Automotive vehicle pricing prediction is important to many different parties who have an interest in the automotive market. This study will demonstrate the importance of collecting all relevant data about the automotive vehicle and pre-processing this data properly using a large dataset with as many characteristics of vehicles as possible and used in combination with various ML algorithms, including Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN).**

Initially, the use of these algorithms in a single classifier method to predict automotive vehicle pricing was found to have some drawbacks. This initial study formed the basis of developing an ensemble (barrier method of combining the best features of the algorithms) method of predicting automotive vehicle pricing. The use of an ensemble method greatly increased the overall accuracy of the prediction's accuracy rate to 92.38%. The research concluded that there are many parameters to consider with respect to trade-offs between computational power and accuracy; the authors believe that the use of an ensemble ML method is a good option to help improve automotive vehicle pricing predictions.

***Index Terms*—Automotive Vehicles, Ensemble Methods, Machine Learning, Pricing Prediction, Random Forest, Support Vector Machine.**

I. INTRODUCTION

Forecasting the cost of a vehicle has been an exciting yet challenging topic. According to figures from the National Automotive Policy Board, there were more than 930,000 registered vehicles in the previous year, with nearly 85% being for personal use (2021). The data show a consistent increase of 2.7% each year and indicate that it is immensely more important than ever to create an

accurate car valuation model (Chang et al., 2023). The rise in vehicle registrations is indicative of the growth in the auto industry and is therefore critically important for several groups, including: consumers trying to buy and sell their own cars, auto dealers, and insurance firms (2023).

Determining a car's market value is complex because TDI and other factors all have an impact on price (Milinkovic, Wetzel, et al., 2020). For example, the car's make and model, age, horsepower, and mileage (Yang, et al., 2022) represent the most common factors used to determine a car's value. Over time, however, the type of fuel used in the car and its fuel efficiency have become important factors (Stylist, et al., 2021) — particularly because of changing fuel prices and a growing public interest in environmental sustainability. Additional factors, such as a car's colour, door count, transmission style, dimensions, safety features, presence of a functioning air conditioning unit (Alhowaity, Ghazal, et al., 2023), and/or the availability of advanced navigation systems, also significantly influence how much a car is worth.

In response to these complications, this paper looks into using advanced machine learning techniques to improve the prediction accuracy of car values. We would like to go beyond using traditional regression-based models and utilize data-driven models that accurately capture the complex relationships of variables that determine the price of automobiles while developing a predictive model capable of adapting to the ever-changing unique characteristics of the automotive marketplace. In performing a comprehensive analysis of various machine learning algorithms, our research intends to provide guidance and methods for improving car price estimation procedures, and as a result, help car buyers and the automotive industry as a whole.

II. RELATED WORK

Researchers have been fascinated with predicting used car pricing, resulting in many studies investigating different computational methodologies. A notable study by Parthian, one of the articles reviewed, states that Support Vector Machines (SVM) outperform traditional multivariate and simple multiple regression models at predicting the prices of leased vehicles. The advantages of SVM lie in the algorithm's ability to be more robust to multi-Dimensional data (more than 1 variable) and ability to resist familiar pitfalls such as over and under-fitting when estimating values. However, this study did not demonstrate how the improvement provided by SVM in regards to common statistical measures (e.g., mean, variance or standard deviation) may have an effect on other aspects related to use car pricing. Thus, further investigation may be warranted.

In his published articles, Deepak has provided an alternative viewpoint on the relationship between the life of a car and its residual value as they relate to hybrid vehicle manufacturers. The use of multiple regression analysis indicates that environmental issues and fuel economy will have an influence on the value of a vehicle, leading to a market preference for hybrid vehicles over traditional vehicles due to their longer life cycles. An innovative methodology was developed by Deepak et al., which involved the use of a neuro-fuzzy knowledge-based system to analyse attributes such as brand, year of manufacture, and type of engine. They also developed the ODAV system, an expert system for optimising the distribution of vehicles at auction, using a regression model based on a k-nearest neighbour's algorithm. This system has achieved great success by

facilitating the exchange of more than 2 million vehicles, based on insights into optimal price and sale location.

A machine learning model for projections related to price of used automobiles was introduced by Gonggie et al., making use of optimal mileage, estimated vehicle life expectancies and the brand of the automobile as predictors (Ramya and Rajeswari, 2023). The advantage of the model was that it was able to account for and model the complex non-linear relationships between variables therefore allowing for much more accurate price predictions compared with previous models based on linear regression.

Samruddhi & Kumar applied machine learning techniques (k-nearest neighbors (k-NN), multiple linear regression (MLR), decision trees (DT), naive Bayes (NB)) to estimate the price of automobiles sold in Mauritius. Even though this was an innovative method to estimate & predict vehicle costs – the experimentation encountered issues with how the DT and NB algorithms process numeric data and the small size of the dataset which impeded the ability to classify accurately.

There are many types of methods used to predict used vehicle values and while each provides useful insights into price prediction, they may also serve to highlight how each prediction concept may be improved by using alternative or combined machine learning techniques to predict used vehicle prices more precisely than may otherwise be done using one single technique or algorithm. Furthermore, most of the studies identified above have used only one machine learning algorithm for vehicle price predictions, so there is evidence to suggest that by using an ensemble approach by combining several of these machine learning methods, we will be able to enhance prediction accuracy and reliability when predicting the price of a used car.

III. DATA AND METHODS

The methodology employed in this study for predicting car prices involves a multi-faceted approach, as illustrated in the conceptual framework (see Fig. 1).

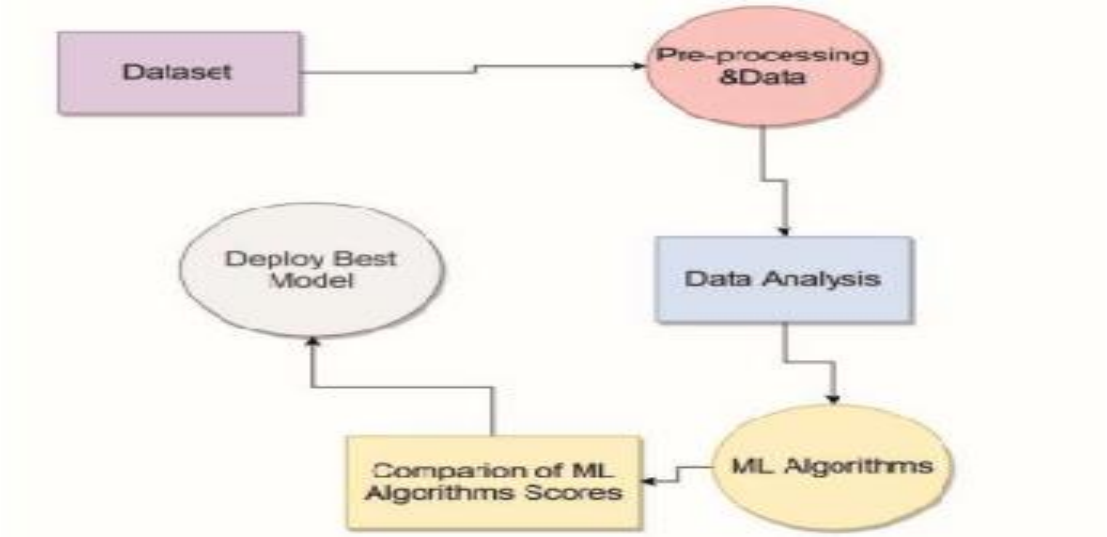

Figure 1. Conceptual Framework for the Car Price Prediction Process

- During our data collection process, we effectively utilized two main resources to provide a diverse and detailed set of data. The first was to obtain additional data through the large Turkish website sahibinden.com, capturing a distinct market climate in those geographic areas at the same time of year as our main source. Combined, this provides a wider view of the car market in these two areas.
- Data collected for each vehicle included many different attributes that were found to be important in predicting price:
- Brand; Model; Condition of vehicle at time of sale; Type of fuel used in vehicle The information collected for the vehicle datasets includes:
  - year of manufacture
  - power (in kW)
  - transmission type
  - mileage
  - colour
  - location (city/state)
  - number of doors
  - Boolean attributes (presence/absence) for various features (e.g., four-wheel drive, navigation, leather seats, etc.)
  - Price (in Turkish Lira) for the dataset.

Web scraping tools (automated) were used to collect data from both websites quickly and easily, as they simulate human behaviour to gather the required data and return it in a structured format. Not only did this save a significant amount of time and effort in data collection, but it also provided accuracy and consistency among the data collected.

After data acquisition, an extensive data pre-processing step was performed, as well as combining the sources into one large dataset. The decision was made to drop the less informative "state" and "city" attributes from both datasets because of redundancy and sparsity, as well as to exclude the "damaged" attribute since it was not consistently reported by either of the original platforms. In total, the resulting dataset included 684 sample observations and provided a more cohesive yet comprehensive format/basis for analysis.

Table 1. Processed Dataset Sample

Brand	Model	Year	Power	Mileage	Fuel Type	Transmission	Number Doors	Four Wheel Drive	Navigation	Leather Seats	Parking Sensors	Price (TL)
Toyota	Corolla	2018	132	32,000	Petrol	Automatic	4	True	False	True	True	800,000
Honda	Civic	2016	158	45,000	Diesel	Manual	4	False	True	False	False	890,000
Ford	Focus	2017	123	37,000	Hybrid	Automatic	4	True	False	True	False	850,000

BMW	3 Series	2015	181	29,000	Diesel	Automatic	4	False	True	True	True	1,590,000
Audi	A4	2019	188	21,000	Petrol	Automatic	4	True	False	True	True	1,650,000
Mercedes-Benz	C-Class	2014	173	55,000	Petrol	Manual	4	False	True	False	True	1,450,000

Note: The boolean attributes (Four Wheel Drive, Navigation, Leather Seats, Parking Sensors) indicate the presence (True) or absence (False) of specific car features. The price is expressed in Turkish Liras (TL).

A python script utilized to pre-process the raw data by removing and cleaning the data and sorting the information to a format that can be used. This script helped eliminate the unfinished records, and normalize the rest of the data into a CSV file format, which could be imported to Math lab, a popular machine learning model development software. This action was essential in the pre-processing stage of the dataset to be used in the following step of applying machine learning methods, to make sure that the input data were of high quality and suitable to predictive analysis.
This study focused on the effectiveness of using one type of machine learning classifier, as has been done in previous studies. But we thought differently and used another batch of classifiers and changed the data division to further test our models. The data collected to conduct this study was split into two subsets, the training (70%) and testing (30%). We built models with the use of Random Forest (RF), Support Vector Machine (SVM) and, most importantly, concentrated on improving the Random Forest classifier in the context of our main analysis.
Random Forest or random decision forest, is an ensemble learning technique that is best applied to both classification and regression problems. RF is a developed algorithm by Ho to reduce the overfitting problem that is usually associated with decision tree algorithms. It works by building a large number of decision trees during training time and returning the mode of the classes (in the case of classification) or mean prediction (in the case of regression) of the individual trees. The advantage of Random Forest is that it can take advantage of large data sets with increased dimensions. It is able to handle thousands of input variables without deleting variables, giving a convenient way of estimating missing data and retaining its accuracy even when a high percentage of the data are missing.

Table 2. Price Classification Based on Price Ranges

From	To	Class
1,000	5,000	100,000–500,000
5,001	10,000	500,001–1,000,000
10,001	15,000	1,000,001–1,500,000
15,001	20,000	1,500,001–2,000,000
20,001	25,000	2,000,001–2,500,000
25,001	30,000	2,500,001–3,000,000

30,001	35,000	3,000,001–3,500,000
35,001	40,000	3,500,001–4,000,000
40,001	45,000	4,000,001–4,500,000
45,001	50,000	4,500,001–5,000,000
50,001	60,000	5,000,001–6,000,000
60,001	70,000	6,000,001–7,000,000
70,001	100,000	7,000,001–10,000,000

Support Vector Machine (SVM) is a very important classification and regression tool. It differentiates categories by creating the greatest possible gap between them. SVM is a binary classifier, which assists in categorizing input data as one of two different categories. The algorithm works best in cases where the data is well labelled and classified and is based on supervised learning.

Techniques. In the case of unlabelled data, unsupervised learning algorithms like Support Vector Clustering (SVC) are advisable, demonstrating the versatility and power of SVM to address a wide range of data conditions.

This change of our methodology to focus on a RF-based approach rather than traditional ANN with a changed data split of 70% training to 30% testing was aimed at exploring the possibility of greater predictive accuracy and generalizability in our dataset. This modification is a result of our intention to deepen our research into the potential of Random Forest in terms of car price prediction, due to its resistance to overfitting and its ability to process high dimensional and complex data sets.

The change in our experimental design, such as the swap of training/testing split, not only aligns with the modern trends in machine learning to obtain a more balanced validation but also enables us to take a critical look at the performance of Random Forest in the competition with other classifiers such as SVM. This is going to guarantee thorough testing of our models so as to understand better their predictive ability and limitations in the context of car price prediction.

Table 3. Single Classifier Approach Accuracy Results

Classifier	Accuracy	Error
RF	85.76%	14.24%
ANN	89.47%	10.53%
SVM	92.38%	7.62%

Table 3 shows the shortcomings of using single machine learning classifiers in order to predict car prices accurately. Considering these results, the current paper suggests an ensemble approach to predicting the prices of cars in a more efficient way. To support this higher-level strategy, we have added a new feature, price rank which can be of three categories, such as cheap, moderate and expensive. Such an adjustment enables a more sophisticated discussion of the prices of cars, not only in terms of numbers.

The ensemble approach capitalizes on the synergy of the three machine learning algorithms that have already been tested as single classifiers: Random Forest (RF), Support Vector Machine

(SVM) and Artificial Neural Network (ANN). Using a combination of these algorithms, we expect to capitalize on the strengths that they have, eliminating the shortcomings of the single classifier mode.

Random Forest algorithm which has the option of meta-estimators was used throughout the entire dataset to classify the cars as cheap, moderate, or expensive. RF works by building several decision tree classifiers on various sub-sets of data. It then averages to improve predictive accuracy and minimize the overfitting risk. The latter approach is especially appropriate to our improved model that features a set of complete characteristics: brand, model, car condition, fuel type, age, power (kilowatts), type of transmission, mileage, color, number of doors, and particular car parameters, such as type of drive, leather seats, navigation system, alarm system, aluminum rims, digital and manual air conditioning, parking sensors, xenon lights, remote unlock, seat heating, pan

Prior to the training of the ensemble model, the numerical attribute "price" was transformed into the nominal classes outlined in Table 4. This change plays a pivotal role in the ensemble approach, and it allows the classifiers to differentiate the established price groups adequately and enhance the accuracy of car price forecasts, as a whole.

Table 4. Nominal Categories of Car Price Attribute

From	To	Class
0	15,000	Budget
15,000	40,000	Mid-Range
40,000	100,000	Premium

IV. CONCLUSION

Car value forecasting is a complex issue, first of all because of the huge number of variables that determine the market value of a vehicle. The paper highlights the paramount significance of careful data gathering and pre-processing as the initial stage in improving the predictive power. By creating python scripts, we were able to normalize, standardize, and clean the dataset, thus minimizing the noise and enhancing the quality of data to be used in machine learning analysis. This sort of pre-processing work is essential to narrow down the dataset but it might not be sufficient to handle the complexities that come with multifaceted datasets, as is the case in the present study.

Early efforts to use a single machine learning algorithm produced accuracy scores lower than 50 percent highlighting the weakness of using one predictive model to do such a complex task. To this end, this paper suggested an ensemble method, which is a combination of many machine learning methods to exploit their synergies. It was found that this strategy led to a significant accuracy increase, with a rate of 92.38, which is much higher than the performance of individual classifier methods.

Nevertheless, one has to consider the trade-offs concerning this sophisticated solution. In particular, ensemble method requires considerably greater computational resources than single

machine learning algorithms. Nevertheless, the disadvantage is compensated by the increased precision of predicted prices of cars obtained using the ensemble method, which presents a bright prospect into the future of research and application in automotive market analysis.

REFERENCES

[1] Alhowaity, A. A. Alatawi, and H. Alsaadi, "Are Used Cars More Sustainable? Price Prediction Based on Linear Regression," Sustainability, vol. 15, no. 2, p. 1640, 2023.

[2] Y. S. Balcioğlu and B. Sezen, "Methodologic Approaches for Transformer Fault Prediction," Recent Advances in Humanities and Social Sciences, p. 191, 2023.

[3] H. Bizimana and A. Altunkaynak, "Investigating the Effects of Bed Roughness on Incipient Motion in Rigid Boundary Channels with Developed Hybrid Gero-Fuzzy versus Neuro-Fuzzy Models," Geotechnical and Geological Engineering, vol. 39, no. 4, pp. 3171–3191, 2021.

[4] L. Chang, M. Mohsin, A. Hasnaoui, and F. Taghizadeh-Hesary, "Exploring Carbon Dioxide Emissions Forecasting in China: A Policy-Oriented Perspective Using Projection Pursuit Regression and Machine Learning Models," Technological Forecasting and Social Change, vol. 197, p. 122872, 2023.

[5] N. A. Deepak, R. Kumar, T. Gupta, S. Gaurav, P. S. Yadav, and B. Pranesh, "Automobile Valuation Prediction Using Machine Learning Based Algorithms," in Proc. 2023 Int. Conf. on the Confluence of Advancements in Robotics, Vision and Interdisciplinary Technology Management (IC-ARVTM), 2023, pp. 1–5.

[6] N. Milanović, M. Milosavljević, S. Benković, D. Starčević, and Ž. Spasenić, "An Acceptance Approach for Novel Technologies in Car Insurance," Sustainability, vol. 12, no. 24, p. 10331, 2020.

[7] S. Parheepan, F. Sanati, and J. Hassan, "Autonomous Unmanned Aerial Vehicles in Bushfire Management: Challenges and Opportunities," Drones, vol. 7, no. 1, p. 47, 2023.

[8] N. Ramya and J. Rajeswari, "A Second User Automotive Value Prediction System for Consumer's Purchasing Using Machine Learning Approach."

[9] K. Samruddhi and R. Kumar, "Used Car Price Prediction Using K-Nearest Neighbor Based Model," International Journal of Innovative Research in Applied Science and Engineering (IJIRASE), vol. 4, no. 3, 2020.

[10] M. Schröder, F. Iwasaki, and H. Kobayashi, "Current Situation of Electric Vehicles in ASEAN," in Promotion of Electromobility in ASEAN: States, Carmakers, and International Production Networks, ERIA Research Project Report FY2021, vol. 3, pp. 1–32, 2021.

[11] J. P. Seybist, S. Busch, R. L. McCormick, J. A. Pihl, D. A. Splitter, M. A. Ratcliff, et al., "What Fuel Properties Enable Higher Thermal Efficiency in Spark-Ignited Engines?" Progress in Energy and Combustion Science, vol. 82, p. 100876, 2021.

[12] Y. Yang, N. Gong, K. Xie, and Q. Liu, "Predicting Gasoline Vehicle Fuel Consumption in Energy and Environmental Impact Based on Machine Learning and Multidimensional Big Data," Energies, vol. 15, no. 5, p. 1602, 2022.